De novo assembling and primary analysis of genome and transcriptome of grey whale *Eschrichtius robustus*

Alexey Moskalev[1,2,*], Anna Kudryavtseva[1], Alexander Grafodatsky[3], Violetta R. Beklemisheva[3], Natalya A. Serdyukova[3], Konstantin V. Krutovsky[4-7], Ivan V. Kulakovsky[1,5,8], Andrey S. Lando[5], Artem S. Kasianov[5], Anastasia Snezhkina[1], Dmitry Kuzmin[6], Julia Putintseva[6], Sergey Feranchuk[9], Mikhail Shaposhnikov[2], Vadim Fraifeld[10], Mitya Toren[10], Vasily Sitnik[8]

[*] Correspondence: amoskalev@list.ru

[1]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991, Russian Federation

[2]Institute of Biology of Komi Science Center of Ural Branch of RAS, Syktyvkar, 167982, Russian Federation

[3]Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, 630090, Russian Federation

[4]Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, 37077 Göttingen, Germany

[5]N.I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, 119333, Russian Federation

[6]Genome Research and Education Center, Siberian Federal University, 660036 Krasnoyarsk, Russian Federation

[7]Department of Ecosystem Science and Management, Texas A&M University, College Station, TX 77843-2138, USA

[8]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow, 143026, Russian Federation

[9]Irkutsk National Research Technical University, 664074 Irkutsk, Russian Federation

[10]Ben-Gurion University of the Negev, Beer-Sheva, 84105, Israel

Abstract

Background

Gray whale *Eschrichtius robustus* is a single member of the family Eschrichtiidae. Eschrichtiidae is considered to be the most primitive, and this species is described as "living fossils".

Results

In this work we for the first time made *de novo* assembling and primary analysis of *E. robustus* genome and transcriptome of kidney and liver. To date, the completeness of the draft genome assembly is about 24%. However, 10895 genes were found due to bioinformatic analysis. Analysis of the transcriptome revealed an increased level of expression of DNA repair and hypoxia-response genes, which is typical for whales.

Conclusions

Further study of the genome and transcriptome of the gray whale will allow us to better understand the ways of the evolution of whales and the mechanisms of their adaptation to deepwater conditions of life.

Keywords

Grey whale, *Eschrichtius robustus*, Genome, Transcriptome

Background

Gray whales, *Eschrichtius robustus* (Lilljeborg, 1861), are the single member of the family Eschrichtiidae. Escrichtidae is one of four families in the suborder Mysticeti (with the Balaenidae, Neobalaenidae and Balaenopteridae). Of these groups, Eschrichtiidae is considered to be the most primitive. Gray whales have been described as "living fossils" because of their short, coarse baleen plates and their lack of a dorsal fin [1].

In this work we for the first time made *de novo* assembling and primary analysis of *E. robustus* genome and transcriptome of kidney and liver.

Methods

Animal samples

The animals were caught by hunters of the indigenous population of Chukotka Autonomous Okrug (Mechigmen bay of the Bering Sea, Lorino) who have permission to hunt this species for food. Tissue biopsies were taken at the time of aboriginal hunting; no animals were killed specifically for this study.

Nucleic acid extraction

Genomic DNA was isolated using phenol-chloroform extraction by standard molecular biology techniques. dsDNA was quantified on the Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA) with the Qubit Broad Range dsDNA kit (Thermo Fisher Scientific, USA), and DNA quality assessment was performed by electrophoresis in 0.6% agarose gel. Only high-quality DNA with greater than 50 kb in size was used for library preparation.

Total RNA was isolated from liver and kidney tissues using RNeasy Mini Kit (Qiagen, Germany) according to the manufacturer's protocol. RNA quantification was performed on the NanoDrop 1000 (NanoDrop Technologies, USA), and RNA integrity number (RIN) was assessed via the Agilent 2100 Bioanalyzer (Agilent Technologies, USA). RNA was further threated with DNase I (Thermo Fisher Scientific, USA) and purified using RNA Clean & Concentrator-5 kit (Zymo Research, USA).

Whole genome sequencing

Three genomic DNA libraries were constructed with fragment sizes 5 Kb and 10 Kb using Nextera Mate Pair Library Prep Kit (Illumina, USA) and insert average size 300 bp with TruSeq DNA Library Prep Kit LT (Illumina, USA) according to the manufacturer's recommendations.

Whole genome sequencing was performed in Genotek (Moscow, Russia) on the Illumina HiSeq 2500 (USA) under the 2 × 100 bp paired-end model.

Transcriptome sequencing

Preparation of cDNA libraries was performed using Illumina TruSeq RNA Sample Preparation Kit v2 (LT protocol) as described [2]. The libraries were sequenced on the Illumina MiSeq System (USA) with corresponding MiSeq Reagent Kit v2 (500 cycle) chemistry. Illumina sequencing was carried out in EIMB RAS “Genome” center (Moscow, Russia).

Genome assembly

The software package CLC Assembly Cell (QIAGEN Bioinformatics, USA) was used for genome assembly. Three types of libraries were used (Table 1).

Genome annotation

The annotation was carried out using a set of software packages and databases (Additional file 1). The primary model for marking the position of genes was obtained by the BUSCO package [3] (Additional file 2). A subset of 3023 groups for Vertebrata was considered. For the detection of genes the AUGUSTUS package [4] with the initial model "human" (*H. sapiens*) was used (Additional file 3). The masking was performed with the RepeatMasker package [5] using RepBase repeats libraries [6] and Dfam [7]. Annotation was carried out with scripts based on the funannotate pipeline [8].

The protein and transcriptomic hints for marking the position of genes were used also. Protein hints were obtained using the Exonerate package [9] (with the appropriate funannotate wrapper) and the protein sequences database SwissProt [10] (for Vertebrata) as well as the protein sequences from the minke whale and bowhead whale assemblies (Additional file 2). Transcriptomic hints were obtained using the blat tool [11] and with the provided transcriptome

assembly. The primary locations of genes obtained using AUGUSTUS was reformatted using the EVidence Modeller package [12] (with the appropriate funannotate wrapper). The finalization of the primary position of genes was carried out using the funannotate pipeline. In total, the primary annotation found 152339 CDS from 43456 parts of genes.

Functional annotation

Search for tRNA genes in genomic sequence was performed with tRNAscan-SE program [13]. The predicted variants with score above 30, not pseudo, and not undetermined were selected to the final annotation. As a result, the final annotation included 2826 predicted tRNAs.

Functional annotation was started by the funnannotate pipeline with disabled annotation by InterPro resource [14, 15]. An annotation was made with the SwissProt protein sequence database [10], Pfam protein families database [16], eggNOG database [17], MEROPS peptidase database [18], and BUSCO families [3]. If protein sequence for the gene was not found in SwissProt, a search for homologues among model mammals in the NCBI Landmark database was conducted.

Then the filtering stage of the marked genes followed. At this stage, only genes with clarified descriptions in SwissProt/NCBI Landmark were selected. One top hit was considered for each marked gene. The total number of unfiltered fragments was 28260, unique hits – 18261, with one hit – 12411. On the average the one hit had 1.5 gene fragments, and fragmented genes were divided into 2.7 parts. The tRNA genes were not filtered.

At filtering stage found genes were selected when more than 30% of the hit from the database were covered by the gene with identity above 60%, and the hit from the database covered more than 60% of the gene. If several genes were found from the database in the same hit, the longest variant was selected. If the top hits for different parts had different IDs (homologues from different organisms), this approach admits annotation of different parts of the same gene, as different genes. Unfortunately, this approach is strongly biased, reduces completeness, does not

allow to reveal duplications, but allowed to follow some limitations on the number and quality of gene marking. After filtering, funannotate pipeline was started again with the annotation by InterPro and GO terms (Table 3, Additional file 4).

Phylogenetics

Phylogenetic tree restoration was performed on the basis of multiple alignment for 322 groups of single-copy orthologous genes, found by the BUSCO methodology, for 16 organisms obtained from the NCBI and Ensembl repositories [19] (Additional file 5). The corresponding protein sequences and CDS for 5152 genes were aligned.

The search for single-copy orthologs was carried out using BUSCO [3]. For the genes represented by several transcripts, only one transcript (with protein product) was selected with the largest BUSCO score. The genes that in all considered genomes have one copy ("complete", in terms of BUSCO) were selected for analysis.

The CDS corresponding to the selected 322 gene groups was aligned using the MAFFT program [20] in the E-INS-i mode, focused on the quality of alignment (with the parameters --ep 0 --genafpair - maxiterate 2000). The resulting alignments were processed by the GBlocks program [21] and glued together into one long sequence. The total length of the sequences for the phylogenetic analysis for CDS was 252,271 base pairs.

The consensus phylogenetic tree was constructed using RAxML [22] with the GTRGAMMAI model. To estimate the convergence of the bootstrapping the autoMRE criterion (extended majority rule consensus tree criterion) was used. The tree of species divergence was constructed by BEAST package [23] with the HKY+Gamma model.The a priori restrictions on divergence times [24] are given in Additional file 6.

Comparison of transcriptome assemblies

In our comparative analysis we used the published whale transcriptome and genome data [25-27]. The details are provided in the Additional file 7. To map transcriptome contigs against genome CDS and Alaska bowhead whale transcriptome we used best hits of blast (executed with default parameters) [28].

Annotation of the obtained gray whale transcriptome assembly and differential gene expression analysis

We used TransDecoder to predict ORFs in assembled contigs and Trinotate [29, 30] to annotate ORFs based on similarity to known orthologous genes. The complete resulting annotation is provided in the Additional file 8, the predicted ORFs are included as an Additional file 9.

To assess gene expression we mapped transcriptome reads of several whale transcriptomes using the gray whale transcriptome assembly as the reference. The reads were trimmed with sickle [31] and cutadapt [32] and mapped using bowtie2 [33] to all contigs carrying ORFs predictions. The mappings in unpaired mode were quite good with nearly 90% of the gray whale reads successfully mapped (80% for minke whale and bowhead whale reads). The mapping in paired mode showed lower but reasonable success rate (70% for gray whale and more than 50% for bowhead and minke whale data). The unpaired mappings were then used for read counting and gene expression analysis to reduce loss of information. The mappings statistics are given in the Additional file 10.

The read counting was performed with HTSeq [34]. Complete read counts are given in the Additional file 11, the distribution of read counts per contig is provided in Additional file 12. Differential expression was assessed with edgeR [35]. One count-per-million expression threshold was used to select the set of reliably expressed transcripts, only 10% of chimeric contigs (with two or more predicted ORFs) passed this expression threshold. The GO enrichment analysis was performed with the Fisher's exact test.

Results and discussion

Draft whole genome sequence assembly and annotation

A whole-genome shotgun sequence approach was used to the genome assembly of the gray whale (*E. robustus*). The liver and kidney transcriptomes were also sequenced and assembled. Approximately 43 Gb (coverage of 17.7×) genome data were generated. The Illumina PE paired-end reads library with reads length 75 bp and Illumina MP mate pair libraries with insert sizes 5 Kb and 10 Kb were sequenced for genome assembly (Table 1). The draft assembly with CLC Assembly Cell (QIAGEN Bioinformatics, USA) software package produced a total of 1779905 scaffolds with an N50 of 10.5 Kb and 2185115 contigs with an N50 of 2.51 Kb (Table 2).

The data of the transcriptome assembly were used for the genome annotation. The primary assessment of genome assembly was carried out using the BUSCO methodology [3]. The number, fragmentation and duplication level of unique orthologs from the different species were evaluated. The genome assemblies of minke whale (*Balaenoptera acutorostrata scammoni*), bowhead whale (*Balaena mysticetus*), and Antarctic minke whale (*Balaenoptera bonaerensis*) were used for comparison (Additional files 2 and 3).

Based on the primary analysis, the expected number of completely reconstituted genes (including duplicated) is about 24%. Apparently, this is due to the relatively small N50 for scaffolds (and small N50 for contigs), comparable (and less, respectively) from the median length for genes in related species (~ 9.3 Kb for minke whale) (Additional file 3).

Known repeats and sequences with low complexity make up about 24.79% of the entire assembly (745.37 Mb) (Table 3, Methods). Despite the fragmentation of the assembly (152339 CDS from 43456 parts of genes were found initially), the use of the filtration procedure, in which contigs with the longest gene fragments were selected (see Methods), allowed to mark 10895 genes (56838 CDS) (Table 3, Additional file 4).

Phylogenetics

Phylogenetic trees were reconstructed on the basis of multiple alignment for 322 groups of single-copy orthologous genes from 16 organisms (Additional file 5). Single-copy "complete" groups were selected in terms of the BUSCO methodology. Figure 1 shows a phylogenetic tree obtained from multiple alignments of examined groups of protein sequences. Despite the insignificant completeness of the genome (about 25% complete by BUSCO, see Additional file 3), the used approach allowed the construction of a fairly plausible tree for groups of protein sequences, keeping the dense of Cetacea cluster. Figure 2 shows a tree of species divergence obtained by multiple alignments of CDS. The used a priori limitations on divergence times [24] are given in Additional file 6. Unfortunately, because of the incompleteness of the draft assembly, there are some deviations in the estimates of the species divergence time from the median estimates given in TimeTree resource [24]. At the same time, the estimated divergence time of *O. orca* and *E. robustus* (34.1, CI: (32.0 - 36.1) MYA) slightly differs from the median time (34.4 CI: (30.6 - 35.5 MYA)) given on the same resource.

The produced gray whale transcriptome assembly provides a better representation of the whale transcriptome compared to previously published data

We have performed compared the gray whale transcriptome assembly (114233 contigs) to the transcriptome assemblies (423657 and 1059024 contigs) and genome CDS annotation (22677 CDSs) of the bowhead whale [25].

The genome CDS annotation contains only nearly 20 thousands of records, which is dramatically different from over a million of transcriptome contigs of the Greenland bowhead whale transcriptome. The total number of contigs of the gray whale transcriptome assembly is ten times smaller and with N50 value being reasonably closer to that of the genomics CDSs. This suggests the produced assembly has less ‘false positive’ and lower number of redundant contigs. To support this statement, we mapped all tested transcriptomes against bowhead whale genome

CDS, as well as Greenland bowhead whale and gray whale transcriptomes against the middle-sized Alaska bowhead whale transcriptome. Indeed, in both tests the mapping showed 2-10 times higher fraction of mapped contigs for the gray whale transcriptome (Additional file 7). Furthermore, the absolute number of reliably mapped contigs and genome CDSs covered by transcriptome contigs' hits were similar for all three tested transcriptome assemblies, which is surprising giving dramatically smaller total size of the gray whale transcriptome assembly. Inter-transcriptome mapping also supports this observation (Additional file 10).

Consistent gene expression across different whale transcriptome samples supports reliability of the genome annotation

To comparatively assess gene expression profiles in kidney and liver of the gray whale we performed standard gene expression analysis using the *de novo* assembled transcriptome as the reference. The gene expression was generally stable in the same organs of different whale species, with the DNA repair and hypoxia-response genes being especially robustly expressed. Next, we performed the gene ontology (GO) enrichment analysis for genes specifically upregulated in the gray whale transcriptome (against minke and bowhead whale data). There were only few genes specifically expressed in gray whale kidney sample and the GO analysis did not show any significant or relevant enrichment. However, GO enrichment of liver data found multiple GO terms enriched (see Additional files 12 and 13), probably linked to the xenobiotic stress response. This might reflect the specific biological state of the studies specimens.

Conclusions

Thus, we made *de novo* assembling and primary analysis of gray whale (*E. robustus*) genome and transcriptome of kidney and liver. According the estimating by BUSCO methodology the completeness of the draft genome assembly is about 24%. After filtration procedure 10895 genes were found. The repeats make up about 24.79% of the entire assembly. The transcriptome

analysis revealed robust expression of DNA repair and hypoxia-response genes, that is consistent with the adaptation of whales to deep diving. The gene ontology enrichment analysis demonstrated upregulation of genes related to xenobiotic stress response in the gray whale liver. This can be due to both the habitat conditions and the physiological state of the individual. Further study of the genome and transcriptome of the gray whale may be useful for understanding the evolution of whales and the mechanisms of adaptation to deepwater conditions.

Abbreviations

bp: base pairs **CDS:** coding DNA sequence **Gb:** gigabase pairs **GO:** gene ontology **Kb:** kilobase pairs **Mb:** megabase pairs **MYA:** million years ago **ORF:** open reading frame **tRNA:** transfer RNA

Declarations

Acknowledgments

Authors are grateful to Michael Zelensky, Alexey Ottoj and The Community of the Chukotka Autonomous Region indigenous “Lorino” (Russian Federation) and PhD S. Blokhin for assistance in the gray whale tissue sample collection. Authors thanks Institute of Molecular and Cellular Biology SB RAS and Irkutsk National Research Technical University for sample providing and Institute of Biology of Komi Science Center of Ural Branch of RAS, Georg-August University of Göttingen, N.I. Vavilov Institute of General Genetics, Siberian Federal University, Texas A&M University, Skolkovo Institute of Science and Technology, Ben-Gurion University of the Negev for help with bioinformatics analysis. Part of this work was performed using EIMB RAS "Genome" center equipment (http://www.eimb.ru/RUSSIAN_NEW/INSTITUTE/ccu_genome_c.php).

Funding

This work was supported by the Russian Science Foundation grant N 14-50-00060.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Data are available at https://www.ncbi.nlm.nih.gov/bioproject/391859

Authors' contributions

AM, MS, AS, KVK, IVK, VS wrote the manuscript text. VRB and NAS carried out DNA extraction. VF, MT, AM, AK carried out the transcriptome assembly. KVK, DK, JP, SF, AM, AK carried out the genome assembly. AM, AK, KVK, IVK, ASL, ASK, AS, DK, JP, SF, MT carried out the bioinformatic analysis. AM, AK, AG, KVK, IVK, VF supervised the bioinformatic research and text of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Not applicable

References

1. Nollman J. The charged border: where whales and humans meet, 1st ed. New York: Henry Holt; 1999.
2. Moskalev A, Shaposhnikov M, Snezhkina A, Kogan V, Plyusnina E, Peregudova D, Melnikova N, Uroshlev L, Mylnikov S, Dmitriev A, Plusnin S, Fedichev P, Kudryavtseva A. Mining gene expression data for pollutants (dioxin, toluene, formaldehyde) and low dose of gamma-irradiation. PLoS ONE. 2014;9:e86051.
3. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210-12.
4. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27:757-63.
5. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org.
6. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6:11.
7. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44:D81-9.
8. Palmer JM. Funannotate: pipeline for genome annotation. http://www.github.com/nextgenusfs/funannotate. 2016.
9. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
10. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204-12.
11. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002;12:656-64.

12. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9:R7.

13. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955-64.

14. Sangrador-Vegas A, Mitchell AL, Chang HY, Yong SY, Finn RD. GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations. Database (Oxford). 2016;2016.

15. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL. InterPro in 2017 - beyond protein family and domain annotations. Nucleic Acids Res. 2017;45:D190-D99.

16. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279-85.

17. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286-93.

18. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2012;40:D343-50.

19. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P. Ensembl 2016. Nucleic Acids Res. 2016;44:D710-6.

20. Katoh K, Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics. 2016;32:1933-42.

21. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56:564-77.

22. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312-3.

23. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969-73.

24. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34:1812-19.

25. Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, Michalak P, Kang L, Bhak J, Yim HS, Grishin NV, Nielsen NH, Heide-Jorgensen MP, Oziolor EM, Matson CW, Church GM, Stuart GW, Patton JC, George JC, Suydam R, Larsen K, Lopez-Otin C, O'Connell MJ, Bickham JW, Thomsen B, de Magalhaes JP. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. 2015;10:112-22.

26. Seim I, Ma S, Zhou X, Gerashchenko MV, Lee SG, Suydam R, George JC, Bickham JW, Gladyshev VN. The transcriptome of the bowhead whale *Balaena mysticetus* reveals adaptations of the longest-lived mammal. Aging (Albany NY). 2014;6:879-99.

27. Yim HS, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, Oh HM, Lee JH, Yang EC, Kwon KK, Kim YJ, Kim TW, Kim W, Jeon JH, Kim SJ, Choi DH, Jho S, Kim HM, Ko J, Kim H, Shin YA, Jung HJ, Zheng Y, Wang Z, Chen Y, Chen M, Jiang A, Li E, Zhang S, Hou H, Kim TH, Yu L, Liu S, Ahn K, Cooper J, Park SG, Hong CP, Jin W, Kim HS, Park C, Lee K, Chun S, Morin PA, O'Brien SJ, Lee H, Kimura J, Moon DY, Manica A, Edwards J, Kim BC, Kim S, Wang J, Bhak J, Lee HS, Lee JH. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 2014;46:88-92.

28. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32:W20-5.

29. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols. 2013;8:1494-512.

30. Das S, Mykles DL. A Comparison of Resources for the Annotation of a *De Novo* Assembled Transcriptome in the Molting Gland (Y-Organ) of the Blackback Land Crab, *Gecarcinus lateralis*. Integr Comp Biol. 2016;56:1103-12.

31. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle. 2011.

32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011;17.

33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357-9.

34. Anders S, Pyl PT, Huber W. HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166-9.

35. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139-40.

Additional files

Additional file 1: The versions of used software packages and databases (PDF 127 kb)

Additional file 2: Assemblies for primary comparison with BUSCO (PDF 110 kb)

Additional file 3: The primary analysis with the BUSCO methodology (PDF 111 kb)

Additional file 4: Functional annotation of genes with funannotate (PDF 8.3 kb)

Additional file 5: Genomic data used for phylogenetic analysis (PDF 29.8 kb)

Additional file 6: A priori estimates of the dates of divergence obtained by using TimeTree resource (PDF 110 kb)

Additional file 7: Comparative assessment of the resulting gray whale transcriptome assembly (PDF 127 kb)

Additional file 8: The complete resulting annotation of the gray whale transcriptome assembly (XLSX 13.1 Mb)

Additional file 9: The predicted ORFs (XLSX 10.3 Mb)

Additional file 10: Transcriptome read mapping statistics (PDF 12.2 kb)

Additional file 11: Complete read counts (XLSX 7 Mb)

Additional file 12: Differential gene expression (PDF 377 kb)

Additional file 13: GO analysis (XLSX 9.79 kb)

Figure legends

Fig. 1 Phylogenetic tree for groups of protein sequences. Phylogenetic tree, built on 322 groups of single-copy orthologous genes. The length of the edges denotes the number of substitutions per site. The bootstrap value for all nodes is 100.

Fig. 2 Tree of species divergence was obtained by multiple alignments for CDS. A priori restrictions on divergence times were used (Additional file 6). The values of the discrepancy time and 95% confidence intervals are shown at nodes.

Tables

Table 1 Libraries sequenced for the genome assembly

Library	Reads length (bp)	Number of reads (pairs)
Illumia PE	75	39011360
Illumina MP (insert size 5 kb)	100	200299976
Illumina MP (insert size10 kb)	100	175370211

Table 2 Main statistics of the genome assembly

Stats for	Total number	N50 (Kb)	Longest (Kb)	Size (Gb)
Contig	2185115	2.51	45.5	2.091
Scaffolds	1779905	10.5	125.01	3.006 (~30% N/X)

Table 3 Statistics of the genome functional annotation

Stats for	Number	Percentage of genome
Repeats	3894603	24.79
Genes (not include tRNA)	10895	2.2255
CDS (not include tRNA)	56,838	0.3461
tRNA	2826	0.0067